멀티모달 대규모 언어 모델 기술 분석

신성한, 용태인, 김재호*
세종대학교
{seonghan.sejong, k62570.sejong}@gmail.com. *kimjh@sejong.ac.kr

A Survey of Multimodal Large Language Model

Shin Seong Han, Yong Tae In, Kim Jae Ho*
Sejong University

요 약

대규모 언어모델의 발전과 함께, 대규모 멀티모달 기반의 언어 모델에 대한 연구가 활발히 진행되고 있다. 텍스트만이 아닌 이미지를 비롯한 여러 형태의 데이터를 통합 처리 하고 고급 추론을 진행하기 위해 모델에 적합한 알고리즘에 대한 논의가 지속되고 있다. 본 논문에서는 대표적인 최신 멀티모달 대규모 언어 모델을 분석하여 각 모델의 독특한 특성을 이해하고자 한다.

Ⅰ. 서 론

최근 인공지능 기술의 발전은 이미지 인식 및 설명 분야에 혁신적인 변화를 가져왔다. 특히, 대규모 언어 모델(LLM, Large Language Model)을 활용한 이미지 설명 모델이 주목을 받고 있다. 이미지 설명 모델들은 이미지 내의 객체, 상황, 감정 등을 인식하고 설명하는데 중요한 역할을 한다. 초기 연구는 단순히 이미지 분류에 초점을 맞추어 진행되었다. 하지만 최근 연구된 모델들은 이미지 속의 사물이나 배경의 복잡한 상호작용이나 사건의 전개를 이해하고 예측하는 능력을 갖추게 되었다. 이를 통해 이미지 한 장으로부터 시각적 이야기를 생성하거나, 이미지 간의 연관성을 찾아내는데 도움을 주고 있다. 이러한 고급 이미지 설명 기술은 인공지능 시스템이 더욱 인간에 가깝게 상황을 인식할 수 있도록 도와주어 로봇 제어나 위치 인식 등 여러 분야에 활용되고 있다. 위와 같은 이미지 설명 모델은 멀티모달 대규모 언어 모델(MLLM, Multimodal Large Language Model)로 불리고 있다.

본 논문에서는 대표적인 MLLM 모델 GPT-4 Vision[1], LLaVA 1.5[2], 그리고 Gemini 1.5 Pro[3]의 공통된 알고리즘적 특징을 검토한다. 그리고 각 모델의 독특한 특성을 비교 분석하여 최신 MLLM 모델들에 대해 깊은 이해를 목표로 한다.

Ⅱ. 본론

2.1 MLLM 기술 분석

GPT-4 Vision, LLaVA 1.5, Gemini 1.5 Pro 모델들은 대규모 데이터 세트를 사용하여 훈련된다. 따라서 높은 정확도를 가지며 준수한 일반화 능력을 가진다. 또한 복잡한 입력을 처리하고 추론을 수행하기 위해 높은 수준의 모델 복잡성을 가지고 있다. GPT-4 Vision의 경우, 최신 기술의 LLM을 사용하여 제한된 모델과 데이터 스케일로 이미지를 분석하는 문제를 해결하여 다중 이미지 및 텍스트를 동시에 입력 받아 해석할 수 있다. GPT-4 Vision은 텍스트와 이미지 같은 다양한 입력을 동시에 처리하기 위해 Interleaved Input Processing 알고리즘을 적용하였다. 이 알고리즘을 통해 입력을 번갈아가며 처리할 수 있도록 하여, 모델은 서로 다른 데이터의 유형을 동시에 이해할 수 있다. 또한 이 알고리즘을 이용하여 모델이 이미지와 텍스트 데이터를 동시에 받아들이고 통합 처리를 진행할 수 있게 한다. GPT-4 Vision에 사용된 주요 메커니즘은 다양한 유형의 데이터를 통합하고 처리하는 멀티모달 퓨전 메커니즘[4]이다. 이 메커니즘에는 3가지 접근 방식이 존재한다. 첫 번째 방식은 조기 융합(Early Fusion)으로, 이 접근 방식을 통해 서로 다른 모달리티 간의 상호작용을 처음부터 고려하여 단일한 표현을 만들 때 사용된다. 두 번째 방식은 후기 융합(Late Fusion)이다. 이 접근 방식을 사용하면 각 모달리티를 개별적으로 처리하고, 최종 결과를 결합하여 각 모달리티의 특징을 독립적으로 분석해 각 데이터 유형의 개별적인 특성을 유지한다. 마지막으로 혼합 융합(Joint Fusion)이다. 이 접근 방식은 초기 단계에서 일부 모달리티는 통합하고, 나머지는 나중에 결합하여 데이터 유형의 특성을 유지하면서, 초기에 상호작용을 고려할 수 있다. GPT-4 Vision은 이 메커니즘을 사용하여 이미지-텍스트 상호작용으로 관계를 이해한다. GPT-4 Vision의 또다른 메커니즘은 키-값 쌍을 사용하는 교차 주의 메커니즘[5]이다. 한 집합의 요소가 쿼리로 작동하여 다른 집합의 요소들이 키와 값으로 작동해 유사성을 측정 및 가중합하여 쿼리에 대한 응답을 생성한다.

다음으로 LLaVA 1.5이다. 이 모델은 Vicuna[6]와 같은 LLM 모델을 이용하여 사용자로부터의 지시 데이터나 질문 데이터를 생성하고, 이 데이터를 기존에 학습된 데이터셋을 사용해 미세 조정하는 Instruction-Tuning[7] 알고리즘을 적용하여 정확도를 높인 모델이다. 특히 시각적 지시에 따라 어떤 특징에 주의를 기울여야 하는지를 학습하는 Visual Instruction-Tuning 알고리즘을 사용한다. 이 알고리즘을 통해 이미지를 분류하거나 객체를 검출하며, 이미지 캡션 생성 등의 작업에서 성능이 향상된다. 또한 이미지를 텍스트로 변환하기 위해 CLIP[8]시각 인코더를 사용한다. CLIP 모델은 Vicuna 모델을 통해 생성된 텍스트와 이미지를 벡터 공간으로 표현한다. 이를 통해 텍스트와 이미지 간의 유사도를 측정하고, 텍스트 설명을 생성하는 과정에서 이미지와 연관된 주요 키워드를 도출한다. 텍스트를 생성하는 Vicuna 모델은 LLaMA[9] LLM 모델을 기반으로 제작된 모델로서 CLIP 모델이 이미지를 텍스트와 유사도가 높은 키워드를 도출해내기 위해 사용한다.

<표 1> MLLM 모델들의 하드웨어 구성 및 기술적 정보

모델	공개 날짜	기반 모델	하드웨어	파라미터	특징
GPT-4 Vision[1]	Sep-2023	SOTA LLM	-	-	• Interleaved Input Processing 알고리즘 • 멀티모달 퓨전 메커니즘
LLaVA 1.5[2]	Oct-2023	Vicuna-13B	8 A100 GPU	13B	• Visual Instruction-Tuning 알고리즘 • CLIP 모델과 Vicuna 모델의 병렬 처리
Gemini 1.5 Pro[3]	Mar-2024	Sparse MoE Transformer	4096 TPUv4	1.6T	• Sparse MoE 알고리즘 • In-Context Learning 기능

마지막으로 Gemini 1.5 Pro[3]는 매우 긴 컨텍스트를 처리하는데 특화되어 있다. 이 모델은 특정 입력에 특화된 파라미터만 활성화하는 Sparse Mixture-of-Expert(MoE) Transformer[10] 기반 모델이다. 기존 MoE[11] 알고리즘은 입력에 대해 제한 없이 파라미터가 활성화되지만, Gemini 1.5 Pro에서 사용하는 Sparse MoE 알고리즘은 활성화되는 파라미터 수를 제한하여 자원의 낭비를 방지할 수 있다. 또한, 각각의 토큰별로 선택되어 훈련을 진행하기 때문에 다양한 도메인에서 뛰어난 확장성을 보여준다. 따라서 여러 문서들이나 몇 시간 길이의 동영상, 매우 긴 오디오를 입력으로 사용할 수 있다. Sparse MoE 알고리즘을 사용하여, 파라미터 개수를 1.6T까지 늘려 대용량의 데이터도 처리할 수 있다. 또한 검색 분야에서는 99% 이상의 recall을 보여주며[3], 긴 문서에서 추가적인 훈련이나 미세 조정 없이 지시에 따라 학습하고 적응하는 In-Context Learning[12] 능력을 보여준다. 추가적으로 <표 1>은 GPT-4 Vision, LLaVA 1.5 및 Gemini 1.5 Pro 모델의 하드웨어 구성과 기술적 특징을 나타낸다.

2.2 MLLM 모델들의 한계점과 향후 연구

<표 1>에 많은 파라미터와 방대한 데이터셋을 바탕으로 훈련된 MLLM 모델들은 높은 계산비용과 접근하기 어렵다는 단점이 있다. 또한 훈련 데이터에 존재하는 편향과 특수한 상황에서의 정확도가 낮고, 왜곡된 결과를 초래할 수 있다. 이를 해결하기 위해 향후 연구로서 모델의 해석 가능성을 높이는 연구와 에너지 효율적 모델링 개선을 제시한다. 모델의 해석 가능성을 높여 AI의 결정 과정을 투명하게 공개하고, 지속적인 윤리적 검토를 통해 기술의 발전이 사회적 가치와 규범에 부합하여야 한다. 또한 에너지 효율성을 개선함으로써 전력 소비를 줄이고 운영 비용을 절감할 수 있다.

Ⅲ. 결론

본 논문에서는 MLLM인 GPT-4 Vision, LLaVA 1.5, 그리고 Gemini 1.5 Pro를 비교하였다. 세 모델들은 전부 대규모 데이터 세트를 사용하여 높은 정확도와 일반화 능력을 보여준다. GPT-4 Vision은 다양한 시나리오에서 세밀하고 정확한 정보를 처리하는 모습을 보여주었고, LLaVA 1.5는 미세 조정 기법을 통해 비교적 적은 데이터셋으로 높은 정확도를 보여주었다. Gemini 1.5 Pro는 긴 입력을 통해 자세한 설명과 풍부한 정보에서 강점을 보여주었다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업(IITP-2024-2021-0-01816)의 연구결과로 수행되었으며, 2024 년도 산업통상자원부 및 산업기술평가관리원 (KEIT) 연구비 지원에 의한 연구임 (RS-2022-00154678)

참 고 문 헌

[1] YANG, Zhengyuan, et al. The dawn of Lmms: preliminary explorations with GPT-4V (Ision). arXiv. arXiv preprint arXiv:2309.17421, 2023.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee. "Improved Baselines with Visual Instruction Tuning" University of Wisconsin-Madison. Microsoft Research, Redmond. pp. 356, Oct. 2023.

[3] REID, Machel, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.

[4] Ngiam, Jiquan, et al. "Multimodal deep learning." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.

[5] VASWANI, Ashish, et al. Attention is all you need. Advances in neural information processing systems, 2017.

[6] PENG, Baolin, et al. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277, 2023.

[7] WANG, Yizhong, et al. Self-instruct: Aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560, 2022.

[8] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[9] TOUVRON, Hugo, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[10] SHAZEER, Noam, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538, 2017.

[11] JORDAN, Michael I.; JACOBS, Robert A. Hierarchical mixtures of experts and the EM algorithm. Neural computation, 1994.

[12] BROWN, Tom, et al. Language models are few-shot learners. Advances in neural information processing systems, 2020.